Thesis/
Reports
Loftsgaarden,
D.O.

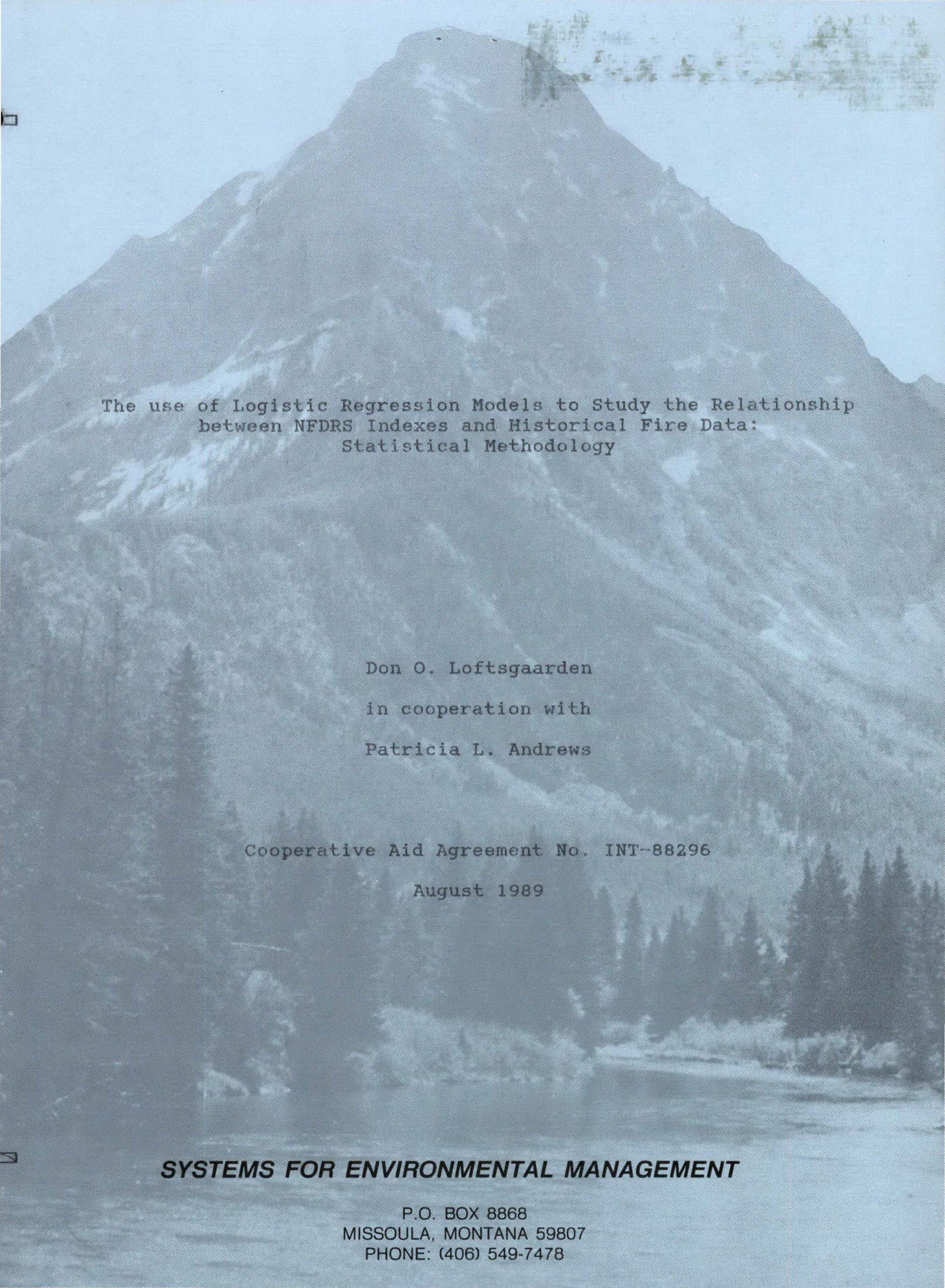
THE USE OF LOGISTIC REGRESSION MODELS TO STUDY THE RELATIONSHIP BETWEEN NFDRS INDEXES AND HISTORICAL FIRE DATA : STATISTICAL METHODOLOGY

BY

DON O. LOFTSGAARDEN

Final Report for Cooperative Agreement #INT-88296-COA with Systems for Environmental Management

FS Contact: Patricia Andrews



The use of Logistic Regression Models to Study the Relationship between NFDRS Indexes and Historical Fire Data: Statistical Methodology

Don O. Loftsgaarden

in cooperation with

Patricia L. Andrews

Cooperative Aid Agreement No. INT-88296

August 1989

SYSTEMS FOR ENVIRONMENTAL MANAGEMENT

P.O. BOX 8868
MISSOULA, MONTANA 59807
PHONE: (406) 549-7478

SYSTEMS FOR ENVIRONMENTAL MANAGEMENT

A Natural Resources Research Group

Systems for Environmental Management (SEM) is a non-profit research and educational corporation based in Missoula, Montana, a regional center for natural resources agencies. Founded in 1977, SEM works cooperatively and under contract with the U.S. Forest Service, National Park Service, Bureau of Land Management, and Bureau of Indian Affairs, as well as state and academic institutions, private individuals and organizations.

SEM's diverse professional staff offers a wide range of research capabilities within the natural resources management field. The full-time staff includes specialists in fire behavior, fire history, recreation management, geography, plant ecology, forestry, meteorology, data analysis, and computer science. SEM also maintains a pool of professional affiliates that can be utilized for projects requiring additional expertise. Areas of recent research emphasis include wilderness recreation management, fire history and ecology, fire planning, and development of computerized resource management tools.

This final report completes the requirements for

Cooperative Agreement No. INT-88296

This report is a *working* draft of an INT General Technical Report.

It should *not* be copied and distributed in this form.

This research was supported in part by funds provided by

the U.S. Department of Agriculture, Forest Service,

Intermountain Research Station.

INTRODUCTION

Logistic regression proved to be a very useful analysis tool in examining the relationship between National Fire Danger Rating (NFDR) indexes and historical fire occurrence data. Specific results of the study will be reported elsewhere. In this paper we use selected examples to describe the statistical methodology. We do this to document the methods we used and also to give guidance to others who might apply these techniques to similar problems. We assume that the reader has some knowledge of statistics. Standard notation and terminology are used.

All readers are familiar with linear regression, the coefficient of determination R^2, and related tests and inferences. Logistic regression is not as widely used and understood. Simple linear regression uses data of the form (x,y) where x is an explanatory (independent) variable and y is a response (dependent) variable which takes values on a continuous scale. Logistic regression uses data of the same form except y is a binary variable having only values 0 or 1.

The following are recent examples uses of logistic regression:

an insect will live or die after a certain dose of spray (Stukel 1988);

a tree will live or die after a fire (Ryan and Reinhardt 19xx);

a lightning strike will either start a fire or not (Latham and Schlieter 19xx);

at a given moisture content, a smoldering fire will continue to burn or go out (Hartford 1989).

In the examples of this paper, the response variable is fire-day which has value 1 if one or more fires occur in an area of concern on that day and 0 otherwise. The explanatory variable is a weather variable or a National Fire Danger Rating System index (Deeming and others 1978). Martell and others (1987) have used logistic regression with success in Canada for problems similar to those discussed here.

An example of linear regression is shown in figure 1. The explanatory variable is father's height and the response variable is daughter's height. A line is fit through the data using least squares. Here $R^2 = .630$. (R^2=1.0 if all points fall on the line.) Based on this model, if a father is 70 inches tall, we predict his daughter to be 66.6 inches tall.

An example of logistic regression model is in figure 2. The explanatory variable is a dose of insecticide, the response variable is insect death with values 0 or 1. Note that none of the points actually fall on the line. For an insecticide dose of xx, the fraction of all insects that are expected to die is xx.

Many applications of logistic regression are classical dose/response studies. An experiment is designed so a dose is administered at a few specific levels. The response is then recorded as either a yes or no. Our application differes from this scenario in several of important aspects. Because the independent variable is calculated from daily weather observations, we have no control of the 'dose.' The NFDRS index can range from zero to as high as 200, so it is necessary to group the data into categories. And one part of our study requires comparing 100 logistic models generated from the same set of weather data (20 fuel models x 5 indexes).

Because of the requirement to construct, evaluate, and compare a large number of logistic models, some techniques that have been used successfully elsewhere have been discarded

for our applications. Other more sophisticated and "messier" to use tests of fit have been proposed recently. Most of these are not yet available in statistical computer packages. They seem to be more appropriate when examining one model in depth.

The purpose of this paper is to report on and illustrate the techniques that have worked well for us. It is not an exhaustive treatise on all aspects of logistic regression. The emphasis of this paper will be on testing logistic regression models. Techniques for one variable and multi-variable models are given. This is currently a very active area of research in statistics and continued improvements in this methodology can be expected over the next few years (for example, Hosmer and Lemestow 1989).

NFDRS/FIRE OCCURRENCE STUDY

The National Fire Danger Rating System (NFDRS) was designed to give an indication of the level of fire danger for each day. It is based on a daily weather reading and on-site conditions defined for a broad area (Deeming et al. 19xx). A seasonal plot of an NFDRS index is shown in figure xx. Although NFDRS indexes are used in making fire management decisions, very little work has been done in looking at the relationship between NFDRS indexes and quantitative measures of the fire situation (Andrews 1987). The need to examine this relationship was pointed out in the Final Report on Fire Management Policy, May 5, 1989. The task force was commissioned as a result of the severe fire season of 1988, some of the most visable fires being in Yellowstone National Park. "Validation of the relationship between current fire management information system components (i.e., drought index, energy release component, 1000 hour fuel moisture, etc.) with actual fire occurrence, severity and size is needed."

In this study we looked at the relationship between NFDRS indexes and fire occurrence. A day was classified as a fire-day if one or more fires were reported on that day. Otherwise it was a no-fire-day. With an NFDRS index as the explanatory variable, we have a situation that seems tailor-made for logistic regression. We had excellent success with our logistic regression models.

In addition to examining the relationship between indexes and fire occurrence, we had several other specific applications to which we hoped to apply the methods:

(1) Test the performance of NFDRS and help define what NFDRS can do. The NFDRS was not designed to predict the behavior of individual fires. It gives an overall rating of the daily fire potential, a concept that is not easy to define and measure. The logistic regression models described in this paper give probability of a fire-day for a index value. These probability curves can be used to help a fire manager calibrate an index for a specific area and to set decision points.

(2) Develop a method for choosing the appropriate fuel model and index for a particular location. NFDRS offers the fire manager a choice of 20 fuel models to define the area of concern. The choice is most often based on a description in terms of vegetation and on experience in what seems to work best. In addition there are 5 components and indexes from which to choose. The fire manager needs an objective means of choosing the 'best' fuel model and index.

(3) Evaluate proposed improvements to NFDRS. In response to deficiencies pointed out in NFDRS, changes have been proposed to the system (Burgan 1988). We wanted a means to evaluate whether the 88 NFDRS was an improvement over the 78 NFDRS.

Our work with logistic regression models has shown that these models can be very useful in all cases we envisioned above. We are very optimistic that logistic regression models will be successfully applied to a wide variety of questions related to NFDRS.

SIMPLE LINEAR REGRESSION REVIEW

It will be useful to have a few facts about simple linear regression as a starting point in discussing logistic regression. The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, ..., n$$

where β_0, β_1 are unknown parameters, x_i's are known constants, e_i's are independent random variables, and $E(e_i) = 0$, $\text{var}(e_i) = \sigma^2$. (The e_i's are often assumed to be normally distributed.) The data is a random sample of n independent pairs

$$(x_1, y_1), ..., (x_n, y_n)$$

Least squares and maximum likelihood estimators for $\hat{\beta}_0, \hat{\beta}_1$ are found by minimizing

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \qquad (1)$$

and are denoted by $\hat{\beta}_0, \hat{\beta}_1$. Also

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, ..., n$$

It is useful to think of the following two models and residual variation of the Y_i's about each of these models.

Model 1:

$\hat{Y} = \bar{y}$ (x is ignored)

$$SSTOT = \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

Model 2:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$

$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$

Then the sum of squares explained by the regression model, Model 2, is

$SSR = SSTOT - SSE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2.$

The coefficient of determination is defined as

$R^2 = SSR/SSTOT.$

This information is often summarized in an ANOVA (Analysis of Variance) table.

ANOVA

Source	df	SS	MS
Regression	1	SSR	MSR
Residual	n-2	SSE	MSE
Total	n-1	SSTOT	

A fit of model 1 and model 2 to a set of data is shown in figure 3.

Since $\hat{\beta}_0, \hat{\beta}_1$ were obtained by minimizing a sum of squared deviations, it is natural to measure variation about a model using squared deviations, i.e. squared distances from a Y_i to a predicted Y_i as we did for the two models above.

Consequences of the assumptions for a simple linear regression model are:

i) $E(Y \mid x) = \beta_0 + \beta_1 x_i$

ii) $\text{Var}(Y \mid x) = \text{Var}(Y) = \sigma^2$ (a constant that does not depend on x), and

iii) $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ (if normal assumption for ϵ's is made).

A MODEL FOR BINARY DATA

In section 1 several examples were given where the response variable Y is a Bernoulli random variable. A Bernoulli random variable is either 0 or 1 where $P(Y = 1) = \pi$. $E(Y) = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi$ and $\text{var}(Y) = \pi \cdot (1 - \pi)$. Often π depends on one or more explanatory variables. For one variable x, $\pi(x) = E(Y \mid x)$ is the probability that Y=1 as a function of x.

There are numerous problems if one tries to use the simple linear regression model of section 2 for binary data. Some of these problems are

i) $E(Y \mid x) = \pi(x)$ is usually not linear

ii) $\text{Var}(Y \mid x) = \pi(x) \cdot (1 - \pi(x))$ is not constant but depends on x

iii) Y is not a normal random variable and

iv) $\pi(x)$ is a probability and hence must have values between 0 and 1.

One approach that has been used is weighted least squares with simple linear regression. The reason for weighted regression is to deal with the variances which are not constant but depend on x. However, most of the problems mentioned above still remain.

A model that has been used in many recent applications with a great deal of success is a logistic regression model (see earlier references). Here it is assumed that

$$\ell n(\pi(x)/(1 - \pi(x)) = \beta_0 + \beta_1 x$$

Solving this expression for $\pi(x) = E(Y \mid x)$

$$E(Y \mid x) = \pi(x) = exp(\beta_0 + \beta_1 x)/(1 + exp(\beta_0 + \beta_1 x))$$

$$= 1/(1 + exp(-\beta_0 - \beta_1 x)).$$

Plotting $\pi(x)$ against x gives the logistic curve (see figure 2). This curve is bounded below by 0 and above by 1. In practice one has n independent observations (x_1, y_1),..., (x_n, y_n) to use in estimating β_0, β_1. For simple linear regression, if all of the data falls on the curve, $R^2 = 1$, but this can never happen for logistic regression. As illustrated in figure 2, all of the data is either 0 or 1. No observations fall on the curve.

LOGISTIC REGRESSION – ONE EXPLANATORY VARIABLE

Logistic Model

For logistic regression the response variable Y has probability distribution

y	0	1
p(y)	$1-\pi$	π

with $E(Y) = \pi$

or $p(y) = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1$

It is assumed that π is a function of an explanatory variable x, i.e., $\pi = \pi(x)$. Thus

$$E(Y \mid x) = \pi(x).$$

As shown above, when $\ell n(\pi(x)/(1-\pi(x)))$ is modeled as a linear function, $\beta_0 + \beta_1 x$, solving for $\pi(x)$ gives

$$E(Y \mid x) = \pi(x) = 1/(1 + exp(-\beta_0 - \beta_1 x)).$$

The data consists of n independent pairs (x_1, y_1),..., (x_n, y_n). If $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates for β_0 and β_1, then

$$\hat{\pi}(x) = 1/(1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 x))$$

is the estimated logistic regression curve. For each x, $\hat{\pi}(x)$ is an estimate for the probability that $Y = 1$.

Maximum likelihood estimation for logistic models

In order to see how to extend the notion of sum of squares used for simple linear regression, it is necessary to look briefly at the likelihood function for the logistic model given above.

$$P(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad y_i = 0, 1$$

where $\pi_i = \pi(x_i) = 1/(1+exp(-\beta_0 - \beta_1 x_i))$.

Let $\underline{\pi} = (\pi_1, ..., \pi_n)$ and $\underline{y} = (y_1, ..., y_n)$.

Then

likelihood function $= L = f(\underline{y} \mid \underline{\pi})$

$$= \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

For finding maximum likelihood estimators it is easier to work with ℓnL.

$$\ell nL = \sum_{i=1}^{n}(y_i \ell n\pi_i + (1-y_i)\ell n(1-\pi_i)) \qquad (2)$$

After replacing π_i by $1/(1+exp(-\beta_0 - \beta_1 x_i))$

and simplification

$$\ell nL = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ell n(1+exp(\beta_0 + \beta_1 x_i)). \qquad (3)$$

Also of interest is $-2\ell nL$ from (2),

$$-2\ell nL = \sum_{i=1}^{n} y_i(-2\ell n\pi_i) + \sum_{i=1}^{n}(1-y_i)(-2\ell n(1-\pi_i)). \qquad (4)$$

Maximum likelihood estimates for β_0, β_1 are found by maximizing (3) as a function of β_0, β_1. If these estimates are $\hat{\beta}_0$ and $\hat{\beta}_1$, then $\hat{\pi}_i = \hat{\pi}(x_i) = 1/(1+exp(-\hat{\beta}_0 - \hat{\beta}_1 x_i))$. A measure of residual variation is found by substituting $\hat{\pi}_i$ for π_i in (4). This will be examined in detail in the following section.

Measuring variability in the logistic situation

If $0 < \pi < 1$ and y is 0 or 1, two measures of "closeness" or "distance between" π and y are now defined and examined.

Definition:

$$S_1(y,\pi) = \begin{cases} -2\ell n\pi & y = 1 \\ -2\ell n(1-\pi) & y = 0 \end{cases}$$

$$= -2[y\ell n\pi + (1-y)\ell n(1-\pi)] \qquad y = 0,1$$

$$= y[-2\ell n\pi + (1-y)(-2\ell n(1-\pi))] \qquad y = 0,1$$

$$S_2(y,\pi) = (y-\pi)^2$$

Numerical examples are given in table 1. When y is near π, both S_1 and S_2 are small; and when y is far from π, both S_1 and S_2 are large. The motivation for S_1 comes from (4).

Consider $y_1, y_2, ..., y_n$ and corresponding $\pi_1, \pi_2, ..., \pi_n$. Recall that the y_i's are 0's and 1's and π_i's are probabilities. Two measures of variations of y's about π's are defined as follows:

$$\sum_{i=1}^{n} S_1(y_i,\pi_i) = -2\sum_{i=1}^{n}[y_i\ell n\pi_i + (1-y_i)\ell n(1-\pi_i)], \qquad (5)$$

$$\sum_{i=1}^{n} S_2(y_i,\pi_i) = \sum_{i=1}^{n}(y_i-\pi_i)^2. \qquad (6)$$

Note that expressions (4) and (5) are the same, i.e., both are $-2\ell nL$. It is $-2\ell nL$ that is minimized to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, so it is the natural measure of variation to use in measuring variation of y_i's about π_i's. In the remainder of this paper (5) is used almost exclusively. It is analogous to the sum of squares for simple linear regression.

There are two sums on the righthand side of (5). The terms in the first sum are non-zero only for y_i's that are 1 and the terms in the second sum are non-zero only for y_i's that are 0.

Formula (5) with $\hat{\pi}_i = 1/(1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 x))$ replacing π_i is equivalent to $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ for simple linear regression.

Closely related to S_2 and (6) is the idea of quadratic score for an observation and average score for a set of observations. This will be discussed further later in this paper.

Analysis of Variance and Tests of Fit for Logistic Regression

Sums of squares

Using (5) as a measure of variation of y_i's about a set of π_i's, it is possible to define "sums of squares" in a fashion similar to simple linear regression.

Starting with $\ell n(\pi/(1-\pi)) = \beta_0$ (i.e., no x in the model) and independent observations $y_1, y_2, ..., y_n$, the maximum likelihood estimator for β_0 is

$\hat{\beta}_0 = \ell n(\overline{y}/(1-\overline{y}))$.

Equating this with $\ell n(\hat{\pi}/(1-\hat{\pi}))$ gives

$\hat{\pi} = \overline{y} = (\sum_1^n y_i)/n = (\#1\text{'s})/n =$ overall fraction of 1's.

The residual variation of $y_1, y_2, \ldots, y_n$ about the constant $\hat{\pi}$ is analogous to the total sum of squares for simple linear regression.

Definition

$$SSTOT = \sum_{i=1}^{n} S_1(y_i, \hat{\pi}) = -2[(\sum_1^n y_i)\ell n\hat{\pi} + (\sum_1^n (1-y_i))\ell n(1-\hat{\pi})]$$

This simplifies to $SSTOT = -2[(\#1\text{'s})\ell n\hat{\pi} + (\#0\text{'s})\ell n(1-\hat{\pi})]$.

Example

If $n = 200$, #1's = 80, #0's = 120, $\hat{\pi} = 80/200 = .4$ and $1-\hat{\pi} = .6$. Then $SSTOT = -2[80\ell n(.4) + 120\ell n(.6)] = 146.61 + 122.60 = 269.21$. This model is shown in figure 4.

In general we have a model $\hat{\pi}(x)$ which depends on x and is based on n independent pairs $(x_1, y_1), \ldots (x_n, y_n)$. For each x we compute $\hat{\pi}_i = \hat{\pi}(x_i)$. A measure of residual variation of y_i's about $\hat{\pi}_i$'s is defined as follows.

Definition

$$SSE = \sum_1^n S_1(y_i, \hat{\pi}_i) = -2[\sum_1^n y_i \ell n \hat{\pi}_i + \sum_1^n (1 - y_i) \ell n (1 - \hat{\pi}_i)]$$

Likewise, we define SS explained = SSR = SSTOT - SSE and $R^2 = SSR/SSTOT$.

CLASSICAL LOGISTIC REGRESSION

Many of the original applications of logistic regression were to dose/response problems. For example, X might be the dosage of an insect spray and Y the response, (Y=1 insect dead, Y=0 insect alive). Many observations are made at a small number of distinct dosages. Theoretically, as dosage increases the fraction of insects killed increases. An example from Stukel (1988) is discussed below.

In such problems, two models are considered.

Model 1: A logistic model

Model 2: A best $\hat{\pi}$ model which defined as the fraction of 1's at each distinct value of x.

We denote these two models by $\hat{\pi}_1(x)$ and $\hat{\pi}_2(x)$ respectively. For each model we compute SSE and SSR and call these SSE_1 and SSR_1 for model 1 and SSE_2 and SSR_2 for model 2. $SSTOT$ does not depend on which model is used.

Shortly we will consider ways to compare these two models but before we do this we make

some observations.

(1) Model 1 has two parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, estimated using all of the data. Model 2 has one parameter for each distinct value of x and each parameter is estimated using only the data for that particular value of x.

(2) The parameters for each model are estimated by minimizing SSE_1 or SSE_2. It turns out that $SSE_2 < SSE_1$ and $SSR_2 > SSR_1$. Of course, $SSTOT = SSE_1 + SSR_1 = SSE_2 + SSR_2$.

(3) If x has only a small number of distinct values and there are several observations at each value of x, then the best $\hat{\pi}$ model (model 2) generally behaves as expected, (i.e., it is monotonically increasing or decreasing with x). In such cases, it is reasonable to compare model 1 with model 2 as a test of how well model 1 fits the data. If, on the other hand, x has many distinct values (perhaps as many as n), and there are only a few observations at each distinct value of x (often only 1), then model 2 can bounce up and down erratically. In this case, comparing model 1 to model 2 is uninformative concerning the fit of model 1 to the data.

(4) Model 2 is not being considered as a potential model but is only used as a summary of the data to have something to compare model 1 against. Model 2 has too many parameters and it is not smoothed at all since only data at each distinct value of x is used in estimating the parameter there. Model 2 does fit the data perfectly in terms of fraction of 1's at each value of x.

SSR_1 is the explained variation for the logistic model and it has 1 degree of freedom. SSR_2 is the explained variation term for model 2 and it has k-1 degrees of freedom where k is

the number of distinct values x assumes. $SSR_2 - SSR_1$ is the extra variation explained by model 2 over what the logistic model explains and this difference has an approximate chi-square distribution with (k-1) - 1 = k-2 degrees of freedom if there is no significant difference between model 1 and model 2. This information is summarized in an analysis of variance table (ANOVA).

ANOVA

Source	df	SS
Logistic model	1	SSR_1
Improvement-best $\hat{\pi}$ over logistic	k-2	$SSR_2 - SSR_1$
Residual	n-k	SSE_2
Total	n-1	SSTOT

Example 1

This example illustrates the use of logistic regression in a dose/response situation. The data comes from Stukel (1988) and is given in table 3, columns 1-3, 5, 7. The treatment is CS_2 applied to beetles. Here Y = 0 if a beetle survives and Y = 1 if a beetle dies. The explanatory variable is x = dosage of CS_2 and it has only 8 values. Summary statistics and the logistic model follow:

$n = 481$, # killed = 291, # surviving = 190, $\hat{\pi} = 291/481 = .605$.

$SSTOT = -2[291\ell n(291/481) + 190\ell n(190/481)] = 645.44$.

$\pi(x)$ = probability that an exposed beetle dies as a function of x.

$\hat{\pi}(x) = 1/(1 + exp(60.7175 - 34.27x))$.

There are eight distinct values $\hat{\pi}(x)$ corresponding to the eight dosages. These values are .0586, .1639, .3620, .6052, .7951, .9032, .9552, .9790.

For the logistic model we find

$$SSE_1 = \sum_{i=1}^{481} S_1(y_i, \hat{\pi}_i) = -2[6\ell n(.0586) + 53\ell n(.9414) + \cdots + 60\ell n(.9790) + 0\ell n(.0210)]$$

$$= 372.47$$

$SSR_1 = SSTOT - SSE_1 = 272.97$ and $R^2 = 272.97/645.44 = .4229$

For the best $\hat{\pi}$ model, which is just the fraction of beetles actually killed at each dosage, we find

$$SSE_2 = -2[6\ell n(6/59) + 53\ell n(53/59) + \cdots + 60\ell n(60/60) + 0] = 361.24$$

$SSR_2 = SSTOT - SSE_2 = 284.20$

Improvement - best $\hat{\pi}$ over logistic $= SSR_2 - SSR_1 = 284.20 - 272.97 = 11.23$

The ANOVA table below summarizes these results:

ANOVA

Source	df	SS	p-value
Logistic model	1	272.97	.0000 *
Improvement-best $\hat{\pi}$ over logistic	6	11.23	.0815 **
Residual	473	361.24	
Total	480	645.44	

* from chi-square with 1 df

** from chi-square with 6 df

The logistic model gives a very reasonable fit to the data. More information is summarized in table 3. The observed columns are the best $\hat{\pi}$ model. Plots of these two models are found in figure 5. Also shown in the figure are the number of 0's and 1's at each value of x. The logistic model has only two parameters and can be used for dosages other than the eight in the data set.

Further explanation of table 3 is in order. The columns marked O (for observed) come from the raw data and need no further explanation. The values in the fraction killed column under E (for expected) are found by substituting dosage values into the logistic model $\hat{\pi}(x)$. To obtain the values in the number killed column under E, the values in the last column in the table are multiplied by the values in the number exposed column. The values in the number surviving column under E are found by subtracting values in column 4 from those in column 2.

A chi-square test of fit statistic can be calculated using observed and expected numbers of 0's and 1's as given in columns 3-6 of table 3. This statistic has an approximate chi-square distribution with k-3 degrees of freedom where k is the number of distinct values for x. This statistic is discussed in detail in a later section.

For the beetle example discussed earlier, we compute this statistic which has approximately a chi-square distribution with 5 df.

$$\chi^2 = (6 - 3.46)^2/3.46 + (13 - 9.83)^2/9.83 + \cdots + (0 - 1.26)^2/1.26 = 10.02$$

P-value = .0745.

The conclusion is the same as that reached earlier.

More examples and further discussion of the chi-square test of fit.

Example 2.

The data for this example come from n = 2501 days during the fire seasons from 1970-1985 in the Lolo National Forest in Montana. For this example x = Energy Release Component (ERC) based on fuel model G. We will denote x as ERC(G). It is an NFDRS index. The dependent variable Y is 1 if one or more fires are discovered on a day and 0 otherwise.

In order to make this example analogous to example 1, the data has been grouped into 12 sets using intervals of ERC(G). The value of x = ERC(G) for all observations in one of these intervals is taken as the midpoint of the interval. These data are summarized in columns 1-4, 6, and 8 of table 4. Some summary results and the logistic model follow.

n =2501, #1's=899, #0's=1602

$$SSTOT = -2[899\ell n(899/2501) + 1602\ell n(1602/2501)] = 3266.83$$

Logistic model:

$$\hat{\pi}(x) = 1/(1 + exp(-.09705x + 3.790))$$

$$SSE_1 = 2805.76 \quad SSR_1 = 461.07 \quad R^2 = .1411$$

Best $\hat{\pi}$ model:

$$SSE_2 = 2791.63 \quad SSR_2 = 475.20$$

Improvement - best $\hat{\pi}$ over logistic = 14.13.

ANOVA

Source	df	SS	p-value
Logistic model (grouped data)	1	461.07	.0000
Improvement-best $\hat{\pi}$ over logistic	10	14.13	.1671
Residual	2489	2791.63	
Total	2500	3266.83	

The logistic model fits the data very well. The additional variation explained by the best $\hat{\pi}$ over the logistic model is minimal.

The chi-square statistic computed using observed and expected fire-days and no-fire-days is 15.33, df=9, and P-value=.0822.

Comments

Because there were many distinct values of ERC(G), the data were grouped by intervals of ERC(G) to get a reasonable best $\hat{\pi}$ model to compare the logistic model against. For this example, the logistic model was also constructed using the grouped data so both models were constructed using the identical data. Generally, one prefers to construct the logistic model using the raw data as information is lost when the data are grouped as above. Example 3 repeats this example using the raw data to construct the logistic model. Grouping of the data to get a best $\hat{\pi}$ model to compare the logistic model against is still necessary. It should be noted that because the raw data logistic model and the best $\hat{\pi}$ model based on the grouped data are constructed from slightly different data, it is no longer true that the best $\hat{\pi}$ model must be the better model in terms of explained sum of squares. Finally, we note that generally there is little difference in the logistic model based on raw data and that based on grouped data.

Example 3

The data and grouping for this example are those of example 2. In this example, however, the logistic model is constructed from the raw data while the best $\hat{\pi}$ model is still based on the grouped data. These data and other information are found in table 5. Summary results and the logistic model follow.

$n = 2501$ #1's=899 #0's=1602

$SSTOT = 3266.83$

Logistic model:

$\hat{\pi}(x) = 1/(1 + exp(-.0999x + 3.8902))$

$SSE_1 = 2809.02$ $SSR_1 = 457.81$ $R^2 = .1401$

Best $\hat{\pi}$ model:

$SSE_2 = 2791.63$ $SSR_2 = 475.20$

Improvement - best $\hat{\pi}$ over logistic = 17.36.

ANOVA

Source	df	SS	p-value
Logistic model (raw data)	1	457.81	.0000
Improvement-best $\hat{\pi}$ over logistic	10	17.36	.0668
Residual	2489	2791.66	
Total	2500	3266.83	

The logistic model fits the data very well. The best $\hat{\pi}$ model explains very little additional variation over the logistic model. The chi-square statistic computed using observed and expected fire-days and no-fire-days is 16.05, df=9, and P-value=.0659.

Figure 6 contains plots of the logistic model and best $\hat{\pi}$ model along with number of 0's and 1's at the midpoint ERC(G) value for each interval.

Warning

All of the explained sum of squares in the preceding three examples have approximate chi-square distributions with degrees of freedom found in the ANOVA tables. We controlled the best $\hat{\pi}$ model to compare the logistic model against.

When using many statistical packages to do logistic regression, there is a defacto grouping of the data by the distinct values of the explanatory variable. This grouping gives rise to a best $\hat{\pi}$ model and many statistical packages automatically compute a statistic which, in the terminology of this paper, is the improvement of the best $\hat{\pi}$ model over the logistic model. As we mentioned earlier, if there are many distinct values for the explanatory variable and only a few observations at each value, this best $\hat{\pi}$ model can be very poor and worthless as a model to compare the logistic model against as a test of fit. With two or

more explanatory variables, the problems can be even worse.

The chi-square statistic based on observed and expected number of 0's and 1's

We now discuss in more detail the chi-square statistic based on observed and expected number of 0's and 1's that we used in examples 1, 2, and 3. We assume that the data are grouped either by distinct values of the explanatory variables (if there are only a few values) or by intervals of the explanatory variable (the most common situation). It is convenient to think of the data as being summarized in a 2-way table where rows correspond to values of Y (0 or 1) and columns are the intervals (or values) for the explanatory variable. Table 6 shows this layout and introduces the notation to be used. There O_{ij} is the number of observations in row i and column j, $O_{i.}$ is a row total and $O_{.j}$ is a column total. As an example, the two rows in Table 6 correspond to columns 6 and 4, respectively, in table 5.

Expected values can be computed as follows. Pick a column, say 3. There are $O_{.3}$ observations in this column which can be thought of as $O_{.3}$ Bernoulli trials. The probability that $Y = 1$ in this column can be estimated using the logistic model. Call this probability $\hat{\pi}_3$. Then the expected number of 0's and 1's in column 3 are computed as

$E_{03} = (1 - \hat{\pi}_3) \cdot O_{.3}$ and $E_{13} = \hat{\pi}_3 \cdot O_{.3}$ respectively.

A statistic which has an approximate chi-square distribution with k-3 degrees of freedom is

$$\chi^2 = \sum_{i=0}^{1} \sum_{j=1}^{k} [(O_{ij} - E_{ij})^2 / E_{ij}]$$

where k is the number of intervals or number of distinct values of the explanatory variable. Roughly speaking, this statistic furnishes another comparison of the logistic model and the best $\hat{\pi}$ model. The question of whether to use the logistic model based on raw data or on grouped data in computing expected values arises here again. We recommend using the

raw data model.

Example 4

This example illustrates another use for the techniques of this paper. The weather and fire data come from n=3711 days for Black Creek in Mississippi. The National Fire Danger Rating System (NFDRS) fire index to be used is Burning Index based on fuel model C which we denote as BI(C). Using the weather data, we generated two sets of BI(C) values, one based on NFDRS models released in 1978 and the other based on improvements in the NFDRS models in 1988. Our goal is to see if the 1988 improvements in the NFDRS models result in detectable improvements in our logistic models. The approach we use here is to construct and test two logistic models–Model 1 based on 1978 NFDRS models and Model 2 based on 1988 NFDRS models. Once this is done, the two models will be compared to see if the 1988 logistic model is better than the 1978 logistic model. Summary information and results are now given for each model.

Model 1. Based on 1978 NFDRS models.

n=3711 # firedays=924 # no-fire-days=2787

$SSTOT = 4165.39 \quad SSE_1 = 3661.59 \quad SSR_1 = 503.8$

Logistic model:

$\hat{\pi}(x) = 1/(1 + exp(-.0880x + 2.1073))$

The data and further information for model 1 are given in table 7.

ANOVA

Source	df	SS	p-value
Logistic model	1	503.80	.0000
Improvement-best $\hat{\pi}$ over logistic	8	44.38	.0000
Residual	3701	3617.21	
Total	3710	4165.39	

The chi-square based on observed and expected fire-days and no-fire-days is 343.21. Using chi-square with df=7, the P-value is .0000. A large piece of this χ^2 is due to the very last cell where observed no-fire-days is 13 and expected no-fire-days is .51. Nevertheless, even if this cell is not used, χ^2 is still large. The evidence is strong that this logistic model does not fit the data very well. A few very large values of BI(C) have a big influence on the model and results here.

Model 2. Based on 1988 NFDRS models.

The version of the 1988 NFDRS models used in this example had 1-hour fuel moisture = 10-hour fuel moisture. One immediate result of the 1988 changes was to cut the range of BI(C) values from 0-106 (1978) to 0-44 (1988).

n=3711 # firedays=924 # no-fire-days=2787

$SSTOT = 4165.39$ $SSE_2 = 3742.88$ $SSR_2 = 422.51$

Logistic model:

$\hat{\pi}(x) = 1/(1 + exp(-.0981x + 2.1209))$

The data and further information for model 2 are given in table 8.

ANOVA

Source	df	SS	p-value
Logistic model	1	422.51	.0000
Improvement-best $\hat{\pi}$ over logistic	8	3.70	.8831
Residual	3701	3739.18	
Total	3710	4165.39	

For these data χ^2=7.51, df=7, P-value = .3778. This logistic model fits the data very well.

A careful examination of all information presented for models 1 and 2 shows that model 2 (1988) is far better than model 1 (1978).

Analysis and Tests of Fit for Logistic Regression Model with

Two or More Explanatory Variables

The logistic regression model and ANOVA table

If there are r explanatory variables $x_1, x_2, \ldots, x_r$, the model with one variable may be extended as follows:

$$E(Y \mid x_1, x_2, \ldots, x_r) = \pi(x_1, x_2, \ldots, x_r) = 1/(1 + exp(-\beta_0 - \beta_1 x_1 - \ldots \beta_r x_r))$$

or, equivalently,

$$\ell n(\pi/(1-\pi)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r$$

The data consist of n independent $(r+1)$–tuples, $(x_{11}, x_{21}, \ldots, x_{r1}, y_1), \ldots, (x_{1n}, x_{2n} \ldots, x_{rn}, y_n)$. Maximum likelihood estimators for $\beta_0, \beta_1, \ldots, \beta_r$ are denoted $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_r$. The approach used to obtain these maximum likelihood estimators is the same as that used for one variable. All of the expressions, definitions, etc., generalize easily. In particular, the definitions for SSE, SSR, and SSTOT are the same. The only difference is that the logistic regression model is now

$$\hat{\pi}(x_1, x_2, \ldots, x_r) = 1/(1 + exp(-\hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_r x_r))$$

and

$$\hat{\pi}_i = \hat{\pi}(x_{1i}, x_{2i}, \cdots, x_{ri}), \quad i = 1, 2, \cdots, n$$

A logistic model with s variables in it would generate the following ANOVA table.

ANOVA

Source	df	SS
Logistic model	s	SSR
Residual	$n-s-1$	SSE
Total	$n-1$	$SSTOT$

This ANOVA table does not have a best $\hat{\pi}$ model row in it. With two or more variables in the logistic model, it is very messy to do the grouping necessary to have a best $\hat{\pi}$ model. In addition, there will, in general, be so many groups that many will have few or zero observations in them. This makes a best $\hat{\pi}$ model useless to compare a logistic model against. Different tests of fit will be used with two or more variables. Before discussing tests of fit, we will discuss another use for the ANOVA tables that is analogous to an ordinary regression technique. Suppose s variables are in a logistic regression model. Of interest is the improvement in the model (in terms of reduction in SSE) if a new explanatory variable is added to the model. We proceed as follows:

1. Model 1: s explanatory variables.

   Compute $SSTOT, SSR_1$, and SSE_1 for this model.

2. Model 2: one new explanatory variable added to the s in model 1.

   Compute SSR_2 and SSE_2 for this model.

3. The improvement in the model is given by

   $SSR_2 - SSR_1 = SSE_1 - SSE_2$

The ANOVA table is

ANOVA

Source	df	SS
Logistic - s variables	s	SSR_1
Improvement - 1 extra variable	1	$SSR_2 - SSR_1$
Residual	$n-s-2$	SSE_2
Total	$n-1$	$SSTOT$

The sum of squares in row 2 of this ANOVA table has an approximate chi-square distribution with 1 degree of freedom and it can be used to test whether there is a significant improvement in the model with the added variable. One is not limited to adding just one variable at a time. This method is the basis for the stepwise logistic regression procedure in the BMDP statistical package. This is not a test of fit of the logistic model to the data. Example 5 illustrates these ideas.

Example 5

The data is from the same set (Lolo National Forest) as that used in examples 2 and 3. Two explanatory variables will be used in constructing logistic regression models for the probability of a fire day. These variables are x_1=maximum daily temperature and x_2=1,000-hour fuel moisture. They are defined as they are used in the NFDRS. Model 1 will include x_1 only and model 2 will include x_1 and x_2. Summary information for these models follows:

n=2501 SSTOT=3266.83 # fire days=899 # no-fire-days=1602

Model 1: x_1 alone.

$SSE_1 = 2749.49$ $SSR_1 = 517.34$

$\hat{\pi}(x_1) = 1/(1 + exp(-.08921x_1 + 7.3769))$

Model 2: x_1 and x_2.

$SSE_2 = 2630.40 \quad SSR_2 = 636.43$

$\hat{\pi}(x_1, x_2) = 1/(1 + exp(-.62619x_1 + .26543x_2 + .99144))$

Improvement - model 2 over model 1 $= SSR_2 - SSR_1 = 119.09$

ANOVA

Source	df	SS	P-value
x_1 alone	1	517.34	.0000
Improvement when x_2 added	1	119.09	.0000
Residual	2498	2630.40	
Total	2500	3266.83	

There is a very significant improvement in terms of sum of squares explained when x_2=1,000-hour fuel moisture is added to a model with x_1=maximum daily temperature in it. Note that this is not a test of fit for the logistic regression model.

A test of fit for any number of explanatory variables

A test of fit for a logistic model involves a grouping of the observations of some sort. For one independent variable, the distinct values of this variable give one possible grouping. The problems with this grouping, when this variable has many values, were discussed earlier. For two independent variables, the analogous grouping would be based on all possible pairs of values for the two variables. The problems encountered for one variable are compounded for 2 or more variables so this grouping is not viable for use for a test of fit.

With one independent variable, a grouping based on intervals of values for this variable was used. For two independent variables, we could divide the values for each variable into intervals (e.g., 12 intervals each). This would then divide the observations up among 12 x 12 = 144 groups or cells. This almost always results in many cells with few or no observations in them, as well as small expected values in many cells. Again, we discard this method of grouping as being not viable for a test of fit.

Having discarded some methods of grouping, we now discuss another method of grouping which is useful in constructing tests of fit for any number of explanatory variables. Good references are Hosmer and Lemeshow (1980,1989) and Lemeshow and others (1982). Once a logistic model is constructed, a $\hat{\pi}$ value can be attached to each observation and these $\hat{\pi}$ values can then be used to group the observations. There are two ways this can be done.

(1) Order the observations by $\hat{\pi}$ values and put the smallest 10% in one group, the next smallest 10% in a second group, etc., giving 10 groups. Because of ties, each group will usually not contain exactly 10% of the observations. There is nothing magical about 10 groups.

(2) Choose 10 intervals, say $[0,.1],(.1,.2],\cdots,(.9,1]$ and put all observations having a $\hat{\pi}$ value in the same interval in the same group. Nothing is magical about 10 and equal length intervals are not necessary. This method can lead to groups that have few or no observations in them and large differences in number of observations in the groups.

Method (1) is preferred and will be illustrated here. The idea of a test of fit is to do the grouping, count the number of observations in each group to be compared with an expected count computed using the logistic regression model. When there is only one independent variable grouping by $\hat{\pi}$ values or by values of the independent variable are nearly the same.

Once the grouping and tallying of 0's and 1's in each group is completed, they can be summarized in a table. See table 9 where O_{ij} = number in row i and column j. The expected number of 1's in each group is found by summing the $\hat{\pi}$'s for all observations in the group. The expected number of 0's is then found by subtraction. The rationale for computing the expected number of 1's is as follows. Each observation, Y, is a Bernoulli ran-

dom variable with $P(Y = 1) \doteq \hat{\pi}$ where $\hat{\pi}$ comes from the logistic model. Thus $E(Y) \doteq \hat{\pi}$. The number of 1's in a group is simply the sum of the Y_i's in the group and hence the expected number of 1's is found by summing the corresponding $\hat{\pi}_i$'s in each group. The expected number of 0's and 1's is denoted by E_{ij} in table 9. In general, each observation in a group could have a different $\hat{\pi}$ and almost certainly all observations in a group will not have the same $\hat{\pi}$.

The following statistic, χ^2, can be used for a test of fit. It has an approximate chi-square distribution with 8 degrees of freedom.

$$\chi^2 = \sum_{i=0}^{1}\sum_{j=1}^{10}(O_{ij} - E_{ij})^2/E_{ij}.$$

We return briefly to the one variable case before looking at examples. In this case, we grouped by intervals of the explanatory variable. To find the expected number of 1's for each interval, it was implicitly assumed that all observations in an interval had the same $\hat{\pi}$ value–that corresponding to the midpoint of each interval. This works well if observations are spread out fairly uniformly in each interval, but this is not always the case. The method of finding expected values discussed just above can be used in this one variable case, too, and this will be illustrated below.

Examples of tests of fit

Examples 6, 7, 8, and 9 are based on the same data and variables as example 5. The first three of these examples carry out tests of fit for the model with only one explanatory variable x_1 = maximum daily temperature in it. The methods to be used for these examples are:

(1) Grouping by intervals of x_1 and computing expected values as we did earlier in this

paper. (Example 6)

(2) Grouping by intervals of x_1 and computing expected values by summing $\hat{\pi}$'s in each interval as discussed above. (Example 7)

(3) Grouping by $\hat{\pi}$ values and finding expected values by summing $\hat{\pi}$ values in each group. (Example 8)

Finally, the method of example 8 will be applied to the logistic regression model with two variables in it, x_1 = maximum daily temperature and x_2 =1,000-hour fuel moisture. (Example 9)

Example 6

Some summary information can be found in example 5. The logistic regression model contains only x_1 = maximum daily temperature and is $\hat{\pi}(x_1) = 1/(1+exp(-.08921x_1+7.3769))$. Table 10 gives the raw data and expected values (computed as indicated in 1 above). From this table $\chi^2 = 12.33$, df = 7, P-value = .0902. This indicates a reasonable fit of the logistic model to the data.

Example 7

The data, model, etc., are the same as examples 5 and 6. Here the expected values are computed as indicated in (2) above. Table 11 summarizes these expected values. From this table, $\chi^2 = 6.02$, df=7, P-value=.5374. This test gives very strong evidence of a good

fit of the logistic model to the data.

Example 8

The data, model, etc., are the same as examples 5, 6, and 7. Grouping is done by $\hat{\pi}$ values and putting approximately 10 % of the observations in each group. Expected values are computed as indicated in (3) above. Table 12 summarizes these expected values. From this table $\chi^2 = 2.70$, df=8, P-value=.9518. Again, we find a very good fit of the logistic model to the data.

We now make some additional comments on examples 6, 7, and 8. In example 6, 8.24 of the $\chi^2 = 12.33$ comes from one interval, 34-50, where most of the observations are above the midpoint temperature of 42. In example 7, some of this problem is taken care of by the way the expected values are computed. This problem occurs in the first place because of the small expected values for this interval. In example 8, the method of grouping guarantees a fairly large number of observations in each group and consequently, there are no small expected values to blow up the value of χ^2. The method of example 8 is the easiest to implement and it is automatically carried out in the BMDP statistical package. Tables 10, 11, and 12 are displayed here so one can get a feel of what is going on and how well the models fit the data. They, of course, are not necessary to carry out the tests.

Example 9

The data is that of example 5. There are two explanatory variables x_1=maximum daily temperature and x_2=1,000-hour fuel moisture. The logistic model is

$$\hat{\pi}(x_1, x_2) = 1/(1 + exp(-.62619x_1 + .26543x_2 + .99144))$$

The groupingis by $\hat{\pi}$ values and expected values are computed by summing $\hat{\pi}$'s in each group. Table 13 gives the expected values. From this table, $\chi^2 = 5.40$, df=8, P-value=.7143. This model also fits the data very well. The effort required to carry out this test of fit is the same regardless of the number of explanatory variables.

A characteristic of a good logistic regression model

Over the range of values for the explanatory variables in the data set, the values of $\hat{\pi}$ should have a large range. For the models in the earlier examples, the range of $\hat{\pi}$ values is roughly .05 to .95 which is excellent. A model with a range of $\hat{\pi}$ values of say, .10 to .25, may not be very useful. This property of logistic regression models is essential for predicting the probability of a fire day as we have been discussing in this paper. Generally, for a logistic model to have this property, the range of values for the explanatory variables must be fairly large. This point will come up again in the next section where we discuss quadratic scores.

QUADRATIC SCORES - EXAMPLES AND COMMENTS

Quadratic scores

Consider $y_1, y_2, \ldots, y_n$ and corresponding π_i's, $\pi_1, \ldots, \pi_n$. Early in this paper two measures of variation $S_1(y, \pi)$ and $S_2(y, \pi)$ were introduced. At that point we argued that $S_1(y, \pi)$ was the natural one to use and since then it has been used exclusively. We now briefly discuss matters related to $S_2(y, \pi)$. Recall that $S_2(y, \pi) = (y - \pi)^2$. We could have defined various sums of squares using S_2 just as we did for S_1. For example, if a logistic model $\hat{\pi}(x)$ gives $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$ then SSE would be defined as

$$SSE = \sum_{i=1}^{n} S_2(y_i, \hat{\pi}_i) = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2$$

Recall that the y_i's are all either 0 or 1. For this reason SSE, based on either S_1 or S_2, is very sensitive to the location of the data with respect to the logistic curve. If most of the data fall at the lower or upper portions of the curve, most of the y_i's will be near the curve and SSE will be small. If most of the data fall in the central portion of the curve, both 0's and 1's are far from the curve and SSE will be large. This is true even for good fitting logistic models.

Closely related to the sums of squares defined using $S_2(y, \pi)$ is the idea of a quadratic score for an observation and average quadratic score for a set of observations. The latter has been used to measure goodness of logistic models. Here we will argue that average quadratic score is not a useful statistic.

Consider $y_1, y_2, \ldots, y_n$ and corresponding $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$ from a logistic regression model $\hat{\pi}$. For

each pair $(y_i, \hat{\pi}_i)$ a quadratic score is defined as

$$\text{score}_i = 1 - 2(y_i - \hat{\pi}_i)^2.$$

Scores have values between -1 and 1 inclusive. When y_i and $\hat{\pi}_i$ are close together, score_i is near 1 and when they are far apart $score_i$ is near -1. Table 14 contains a few numerical cases.

A measure of how good a logistic model is that is sometimes used is average score defined as follows:

$$\text{Average score} = \sum_{i=1}^{n}[1 - 2(y_i - \hat{\pi}_i)^2]/n$$
$$= \sum_{i=1}^{n}[1 - 2 \cdot S_2(y_i, \hat{\pi}_i)]/n.$$

The average score can be written as

$$[n - 2 \cdot \sum_{i=1}^{n} S_2(y_i, \hat{\pi}_i)]/n = 1 - 2 \cdot SSE/n$$

where SSE is based on S_2.

Notes:

(i) When SSE is large average score is small and when SSE is small, average score is large.

(ii) It is possible to have two different models with the same SSE and hence the same average score, but with one SSE being 10% of SSTOT and the other SSE being 90% of SSTOT.

(iii) As mentioned earlier, SSE and consequently average score are very dependent on where the data fall with respect to the logistic curve. This is the main reason average score is not useful as a statistic for evaluating logistic models.

Examples using quadratic scores

These ideas are illustrated with two examples. Both estimate the probability of a fire-day as a function of x = ERC(G). Example 10 is based on data from the Missoula District of the Lolo National Forest. Example 11 does this same thing, except it is based on data considering the entire Lolo National Forest as a unit rather than individual districts. Example 11 is based on a slightly different set of data than that used for a similar example earlier in this paper.

Example 10. Missoula District of Lolo National Forest

This example is based on approximately 15 years of data from the Missoula District of Lolo National Forest. Summary information follows.

n=2442 # fire-days = 294 # fire-days = 2148 $\hat{\pi}$=294/2442 = .1204.

Logistic model:

$$\hat{\pi}(x) = 1/(1 + exp(4.27604 - .07034x)).$$

$$SSTOT = \sum_{i=1}^{2442} S_2(y_i, \hat{\pi}) = \sum_{1}^{2442} (y_i - \hat{\pi})^2$$

$$= 2148(0 - .1204)^2 + 294(1 - .1204)^2$$

$$= 31.13 + 227.47 = 258.60.$$

$$SSE = \sum_{i=1}^{2442} (y_i - \hat{\pi}(x_i))^2 = 243.36 \quad SSR = 15.24 \quad R^2 = .0589.$$

Average score = 1-2·SSE/2442 = .80.

The data and expected values are found in table 15. The expected values were found using the $\hat{\pi}$ values corresponding to the midpoint of each interval of ERC(G). From table 15 we get $\chi^2 = 10.56$, df = 8, p-value = .2279. The logistic model fits these data quite well.

Example 11. Lolo National Forest

This example is based on approximately 15 years of data from the Lolo National Forest. Summary information follows.

n = 2442 # fire-days=831 # no-fire-days=1611 $\hat{\pi}$=831/2442=.3403.

Logistic model: $\hat{\pi}(x) = 1/(1 + exp(-.09110x + 3.48770))$.

SSTOT = 548.22 SSE = 458.55 SSR = 89.67 R^2 = .1636.

Average score = .62.

The data and expected values are found in table 16. The expected values were found using the $\hat{\pi}$ values corresponding to the midpoint of each interval of ERC(G). From table 16 we get χ^2 = 10.70, df = 8, p-value = .2193. This logistic model also fits the data quite well.

Comments on examples 10 and 11.

(1) The influence of SSE on average score is readily seen. In example 10, SSE is 96% of SSTOT, whereas in example 11, SSE is only 84% of SSTOT. However, since SSE = 243.37 in example 10 is much smaller than SSE = 458.55 in example 11, the average score for example 1 is much larger than that for example 2 (.80 versus .62).

(2) This comment tells us essentially the same thing from a different perspective. The majority of the observations for example 10 are near the lower tail of the logistic model resulting in a small SSE and large average score. The range of $\hat{\pi}$ values in example 11 is much larger than that for example 10. Many of the observations fall in the middle portion of the logistic curve resulting in a large SSE and smaller average score.

(3) Examining the data and results summarized above for these two examples shows that both are quite good and that the model for example 11 is perhaps the better of the two. Average scores tell us nothing about the fit of the two models and would seem to say the example 10 model is best.

(4) The recommendations is do not use average scores for tests of fit or comparison of

models.

Example 12. Martel et al. (1987)

As another example illustrating the use of average score and its problems, we draw an example from Martell et al. (1987). In this paper, logistic models were constructed for the probability of a fire day using 17 years of historical data. Models were constructed for people-caused fires (for several causes and five different seasons). These models were field tested by making predictions with them in 1984. They used average score based on their 1984 observations and predictions as an indicator of how good their models were. Typical average scores were .991, .831, .996, .999, 1.0, etc. At first glance, these seem impressive. However, 1984 was a very wet year with only six people-caused fires.

Consider the early-summer season and people/recreation-caused fires. During this season there were 49 days and 1 fire (1 fire day). The data consist of one 1 and 48 zeroes. Because the year was wet, it is likely that the fire indexes had low values nearly all of the time. This would mean that all of the observations were near the lower tail of the logistic curve where $\hat{\pi}$ is flat and near 0. Because nearly all observations are 0's, SSE is small and average score is very large. Thus these average scores say little or nothing about how good the models are or how well they fit the data. Instead, these average scores say something about where the data fell with respect to the logistic curves. Again, the conclusion is that average score is not a useful statistic to use in evaluating logistic models.

It should also be noted that all of the procedures discussed earlier in this paper are designed to test the models using the data that were used to construct the models in the first place. Generally, this gives one lots of observations to work with. Thus, the problems of few observations and an atypical year as described in the previous paragraph are avoided.

Final Comments

Computations for most of the procedures discussed in this paper can be done using statistical packages such as SAS, BMDP, SPSSX, or GLIM. BMDP has a stepwise logistic regression procedure which is based on the sum of squares defined using S_1 in this paper. BMDP seems easiest to use and gives nearly all of the results discussed here.

The procedures discussed here were chosen so that a large number of logistic models (often 100 or more) could be constructed, evaluated, and compared quickly. We have used the procedures discussed in this paper for this purpose quite successfully.

Other more sophisticated and "messier" to use tests of fit have been proposed recently. Most of these are not yet available in statistical computer packages. They seem to be more appropriate when examining one model in depth. Some of these new procedures can be found in the references to this paper. Tests of fit for logistic regression models is currently a very active area of research in statistics. New and better methods can be expected to appear for some time.

Literature Cited

(A short selection)

Bradshaw, L. S. 1988. Fire business utility programs–user guide. Systems for Environmental Management and Intermountain Fire Sciences Laboratory.

Cox, D. R. 1970. The analysis of binary data. Methuen and Co., Ltd.

Deeming, John E.: Burgan, Rogert E., and Cohen, Jack D. 1978. The National Fire Danger Rating System–1978. USDA Forest Service General Technical Report INT-39.

Efron, Bradley. 1978. Regression and ANOVA with zero-one data: measures of residual variation. J. of Amer. Stat. Assoc. 73: 113-121.

Hosmer, D. W.; Lemeshow, S. 1980. Goodness of fit tests for the multiple logistic regression model. Comm. in Statistics–Theoretical Methods, A9(10):1043-1069.

Hosmer, D. W.; Lemeshow, S. 1989. Applied Logistic Regression. John Wiley and Sons.

Landwehr, J. M.; Pregibon, D.; Shoemaker, A. C. 1984. Graphical methods for assessing logistic regression models. J. of Amer. Stat. Assoc. 79: 61-83.

Lemeshaw, S.; Hosmer, D. W., Jr. 1982. A review of goodness of fit statistics for use in the development of logistic models. Amer. J. of Epidemiology. 115(1): 92-106.

Lemeshaw, S.; Teres, D.; Avrunih, J. S.; Pastides, H. 1988. Predicting the outcome of intensive care unit patients. J. of Amer. Stat. Assoc. 83: 348-356.

Loftsgaarden, D. O. 1988. Logistic models for fire occurrence using fire indexes and weather variability–Mississippi (Black Creek), Mississippi (Bude), and Montana (Lolo). Cooperative Agreement INT-88296, Systems for Environmental Management and Intermountain Fire Sciences Laboratory.

Loftsgaarden, D. O. 1987. The relationship between the National Fire Danger Rating System indexes and fire occurrence–an exploratory study. Cooperative Agreement INT-87228, Systems for Environmental Management and Intermountain Fire Sciences Laboratory.

Martell, D. L.; Otakel, S.; Stocks, B. J. 1987. A logistic model for predicting daily people-caused fire occurrence in Ontario. Can. J. For. Res. 17: 394-401.

Pregibon, D. 1981. Logistic regression diagnostics. The Annals of Statistics. 9(4): 705-724.

Steele, B.; Ryan, K. 1987. Goodness of fit in binary logistic regression. Research Work Unit FS-INT-2108, Intermountain Fire Sciences Laboratory.

Stukel, Therese A. 1988. Generalized logistic models. J. of Amer. Stat. Assoc. 83: 426-431.

Table 1. Numerical values for S_1 and S_2.

π	.1	.1	.4	.4	.5	.5	.75	.75	.95	.95
y	0	1	0	1	0	1	0	1	0	1
$S_1(y,\pi)$	.211	4.605	1.022	1.833	1.386	1.386	2.773	.575	5.991	.103
$S_2(y,\pi)$	.01	.81	.16	.36	.25	.25	.56	.06	.90	.003

Table 2. Beetle data from Stukel paper.

Dosage CS_2	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Number beetles exposed	59	60	62	56	63	59	62	60
Number beetles killed	6	13	18	28	52	53	61	60

Table 2. Summary information for beetle poison example.

(1)	(2)	Number killed (3)	(4)	Number surviving (5)	(6)	Fraction killed (7)	(8)
Dosage	Number exposed	O	E	O	E	O	E
1.6907	59	6	3.46	53	55.54	.1017	.0586
1.7242	60	13	9.83	47	50.17	.2167	.1639
1.7552	62	18	22.44	44	39.56	.2903	.3620
1.7842	56	28	33.89	28	22.11	.5006	.6052
1.8113	63	52	50.09	11	12.91	.8254	.7951
1.8369	59	53	53.29	6	5.71	.8983	.9032
1.8610	62	61	59.22	1	2.78	.9839	.9552
1.8839	60	60	58.74	0	1.26	1.0000	.9790
	481	291		190			

Columns headed O are from observed data
Columns headed E are from logistic model

Table 4. Logistic model using ERC(G) and grouped data.

ERC(G) range	ERC(G) midpoint	Days	Fire-days		No-fire-days		Fraction fire-days	
			O	E	O	E	O	E
0–0	0	11	1	.24	10	10.76	.0909	.0221
1–12	6.5	126	8	5.13	118	120.87	.0635	.0407
13–17	15	119	12	10.51	107	108.49	.1008	.0883
18–26	22	442	71	70.92	371	371.08	.1606	.1604
27–29	28	278	68	70.86	210	207.14	.2446	.2549
30–32	31	306	91	96.09	215	209.91	.2974	.3140
33–35	34	300	96	113.95	204	186.05	.3200	.3798
36–39	37.5	321	156	148.44	165	172.56	.4860	.4624
40–45	42.5	379	240	220.91	139	158.69	.6332	.5829
46–49	47.5	119	78	82.61	41	36.39	.6555	.6942
50–54	52	77	58	59.94	19	17.06	.7532	.7784
55–59	57	23	20	19.57	3	3.43	.8696	.8509
		2501	899		1602			

Table 5. Logistic model using ERC(G) and ungrouped data.

ERC(G) range	ERC(G) midpoint	Days	Fire-days O	Fire-days E	No-fire-days O	No-fire-days E	Fraction fire-days O	Fraction fire-days E
0–0	0	11	1	.22	10	10.78	.0909	.0200
1–12	6.5	126	8	4.74	118	121.26	.0635	.0377
13–17	15	119	12	9.97	107	109.03	.1008	.0838
18–26	22	442	71	68.71	371	373.29	.1606	.1555
27–29	28	278	68	69.79	219	208.21	.2446	.2511
30–32	31	306	91	95.31	215	210.69	.2974	.3115
33–35	34	300	96	113.71	204	186.29	.3200	.3790
36–39	37.5	321	156	148.97	165	172.03	.4860	.4641
40–45	42.5	379	240	222.84	139	156.16	.6332	.5880
46–49	47.5	119	78	83.49	41	35.51	.6555	.7016
50–54	52	77	58	60.57	19	16.43	.7532	.7866
55–59	57	23	20	19.75	3	3.25	.8696	.8586
		2501	899		1602			

Table 6. Layout for chi-square test of fit.

Y	1	2	3	. . .	k	
0	O_{01}	O_{02}	O_{03}		O_{0k}	$O_{0\cdot}$
1	O_{11}	O_{12}	O_{13}		O_{1k}	$O_{1\cdot}$
	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot 3}$		$O_{\cdot k}$	n

Note

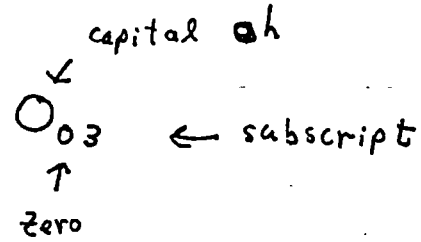


Table 7. Logistic model based on BI(c) and 1978 NFDRS - ungrouped data

BI(c) range	BI(c) midpoint	Days	Fire-days O	Fire-days E	No-fire-days O	No-fire-days E	Fraction fire-days O	Fraction fire-days E
0	0	479	34	51.92	445	427.08	.0710	.1084
1-3	2	557	72	70.52	485	486.48	.1293	.1266
4	4	580	101	85.48	479	494.52	.1741	.1474
5	5	353	58	56.05	295	296.95	.1643	.1588
6-10	8	450	76	88.78	374	361.22	.1689	.1973
11-16	13.5	372	94	106.06	278	265.94	.2527	.2851
17-24	20.5	553	271	234.89	282	318.11	.4901	.4248
25-29	27	200	120	113.35	80	86.65	.6000	.5668
30-36	33	133	77	91.67	56	41.33	.5789	.6893
37-106	71.5	34	21	33.49	13	.51	.6176	.9850
		3711	924		2787			

Table 8. Logistic model based on BI(c) and 1988 NFDRS-ungrouped data.

BI(c) range	BI(c) midpoint	Days	Fire-days O	Fire-days E	No-fire-days O	No-fire-days E	Fraction fire-days O	Fraction fire-days E
0	0	714	75	76.46	639	637.54	.1050	.1071
1-3	2	385	42	49.03	343	335.97	.1091	.1273
4-5	4.5	502	88	78.90	414	423.10	.1753	.1572
6-7	6.5	299	51	55.30	248	243.70	.1706	.1849
8-10	9	388	83	87.22	305	300.78	.2139	.2248
11-14	12.5	512	150	148.56	362	363.44	.2930	.2902
15-20	17.5	549	227	219.77	322	329.23	.4135	.4003
21-24	22.5	190	100	99.10	90	90.90	.5263	.5216
25-32	28.5	145	88	96.08	57	48.92	.6069	.6626
33-44	38.5	27	20	22.67	7	4.33	.7407	.8397
		3711	924		2787			

Table 9. Layout of observed and expected values.

	Group									
Y	1	2	3	4	5	6	7	8	9	10
0	O_{01}/E_{01}	O_{02}/E_{02}	.	.	.					$O_{0,10}/E_{0,10}$
1	O_{11}/E_{11}	O_{12}/E_{12}	.	.	.					$O_{1,10}/E_{1,10}$

Table 10. Method 1: grouping by values of MAXTEMP and computing expected values using midpoint of each MAXTEMP interval.

MAXTEMP range	MAXTEMP midpoint	Days	Fire-days O	Fire-days E	No-fire-days O	No-fire-days E	Fraction fire-days O	Fraction fire-days E
34-50	42	96	7	2.48	89	93.52	.0729	.0258
51-59	55	274	23	21.37	251	252.63	.0839	.0780
60-67	63.5	387	53	59.17	334	327.83	.1370	.1529
68-74	71	409	109	106.59	300	302.41	.2665	.2606
75-78	76.5	310	116	113.24	194	196.76	.3742	.3653
79-83	81	355	166	164.12	189	190.88	.4676	.4623
84-87	85.5	276	158	155.19	118	120.81	.5725	.5623
88-91	89.5	194	116	125.58	78	68.42	.5979	.6473
91-94	92.5	128	94	90.34	34	37.66	.7344	.7058
95-102	98.5	72	57	57.87	15	14.13	.7917	.8038
		2501	899		1602			

Table II. Method 2: grouping by values of MAXTEMP and computing expected values by summing $\hat{\pi}$'s for observations in each interval.

MAXTEMP		Fire-days		No-fire-days	
range	Days	O	E	O	E
34-50	96	7	3.94	89	92.06
51-59	274	23	23.02	251	250.98
60-67	387	53	60.37	334	326.63
68-74	409	109	107.98	306	301.02
75-78	310	116	114.39	194	195.61
79-83	355	166	164.37	189	190.63
84-87	276	158	153.73	118	122.27
88-91	194	116	124.55	78	69.45
91-94	128	94	90.75	34	37.25
95-102	72	57	55.94	15	16.06
	2501	899		1602	

Table 12. Method 3: grouping by values of $\hat{\pi}$ and computing expected values by summing $\hat{\pi}$'s for observations in each interval.

		Fire-days		No-fire-days	
$\hat{\pi}$ interval	Days	O	E	O	E
[0, .085]	258	18	15.48	240	242.52
(.085, .140]	247	27	28.41	220	218.59
(.140, .200]	252	38	43.34	214	208.66
(.200, .270]	220	57	51.92	163	168.08
(.270, .350]	249	77	76.19	172	172.81
(.350, .400]	250	91	94.25	159	155.75
(.400, .490]	292	133	132.28	159	159.72
(.490, .560]	222	118	117.44	104	104.56
(.560, .660]	287	173	176.51	114	110.49
(.660, .850]	224	167	163.07	57	60.93
	2501	899		1602	

Table 13. Logistic model with two variables; computing expected values by summing $\hat{\pi}$'s for observations in each interval.

		Fire-days		No-fire-days	
$\hat{\pi}$ interval	Days	O	E	O	E
[0, .068]	249	11	10.21	238	238.79
(.068, .129]	250	24	24.00	226	226.00
(.129, .191]	252	33	40.07	219	211.93
(.191, .256]	249	54	55.28	195	193.72
(.256, .322]	252	81	73.33	171	178.67
(.322, .410]	242	85	88.09	157	153.91
(.410, .505]	254	113	114.81	141	139.19
(.505, .595]	250	148	137.75	102	112.25
(.595, .700]	249	160	159.61	89	89.39
(.700, .890]	254	190	195.83	64	58.17
	2501	899		1602	

Table 14. Numerical values for scores.

y	1	0	1	1	0	0	1	1	0
$\hat{\pi}$	1	0	.9	.5	.5	.9	.1	0	1
$re = 1-2(y-\hat{\pi})^2$	1	1	.98	.5	.5	$-.62$	$-.62$	-1	-1

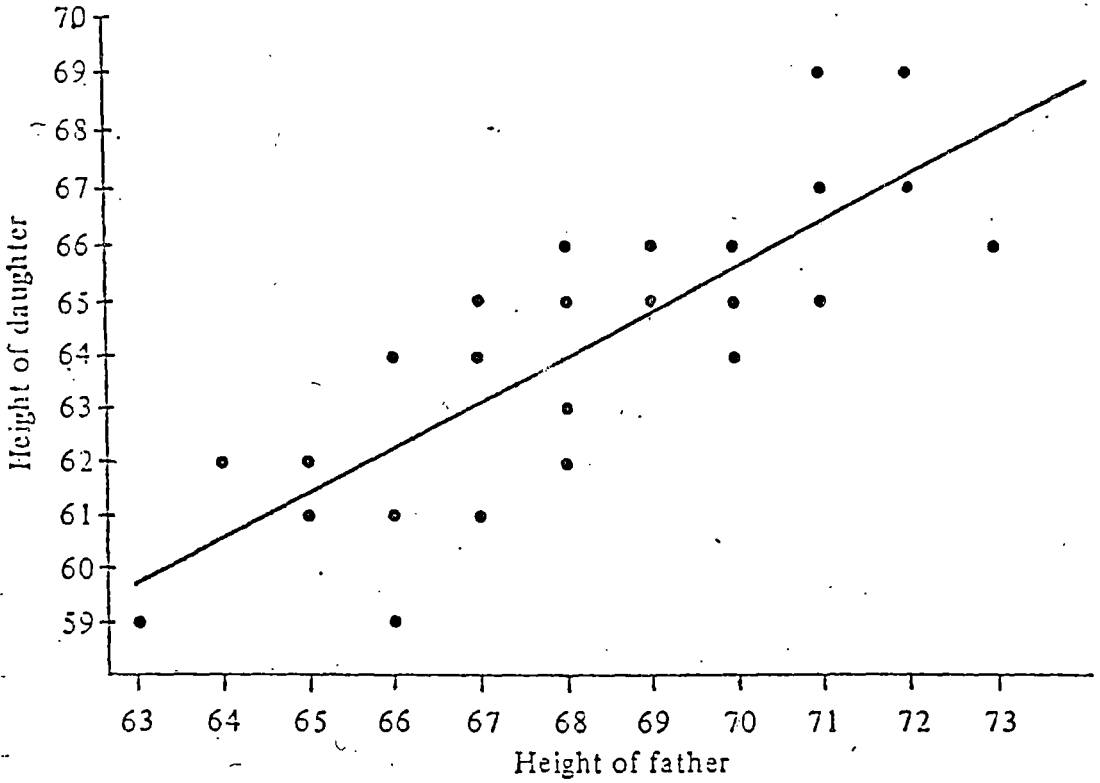


Figure 4. Heights of fathers and daughters.

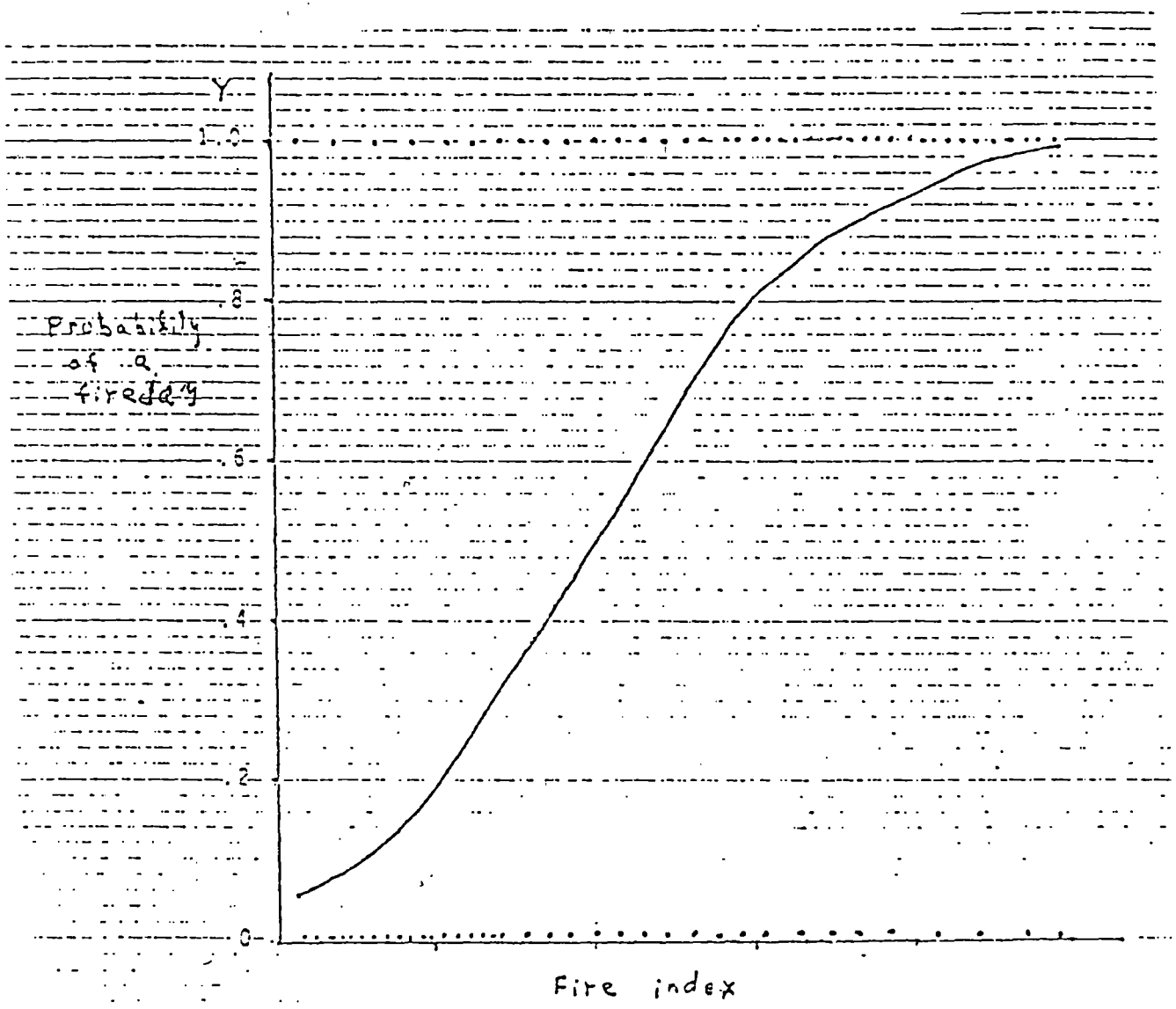


Figure 2. A set of binary data and a logistic curve

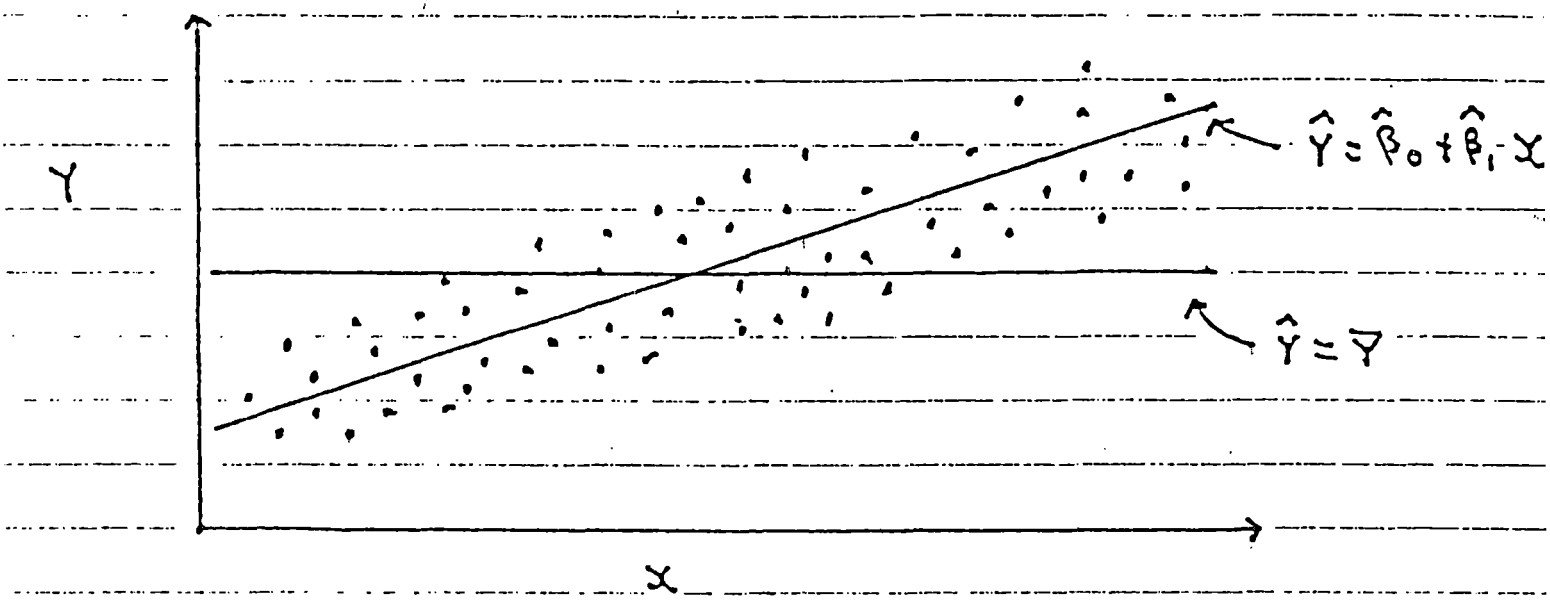


Figure 3. Simple linear regression.

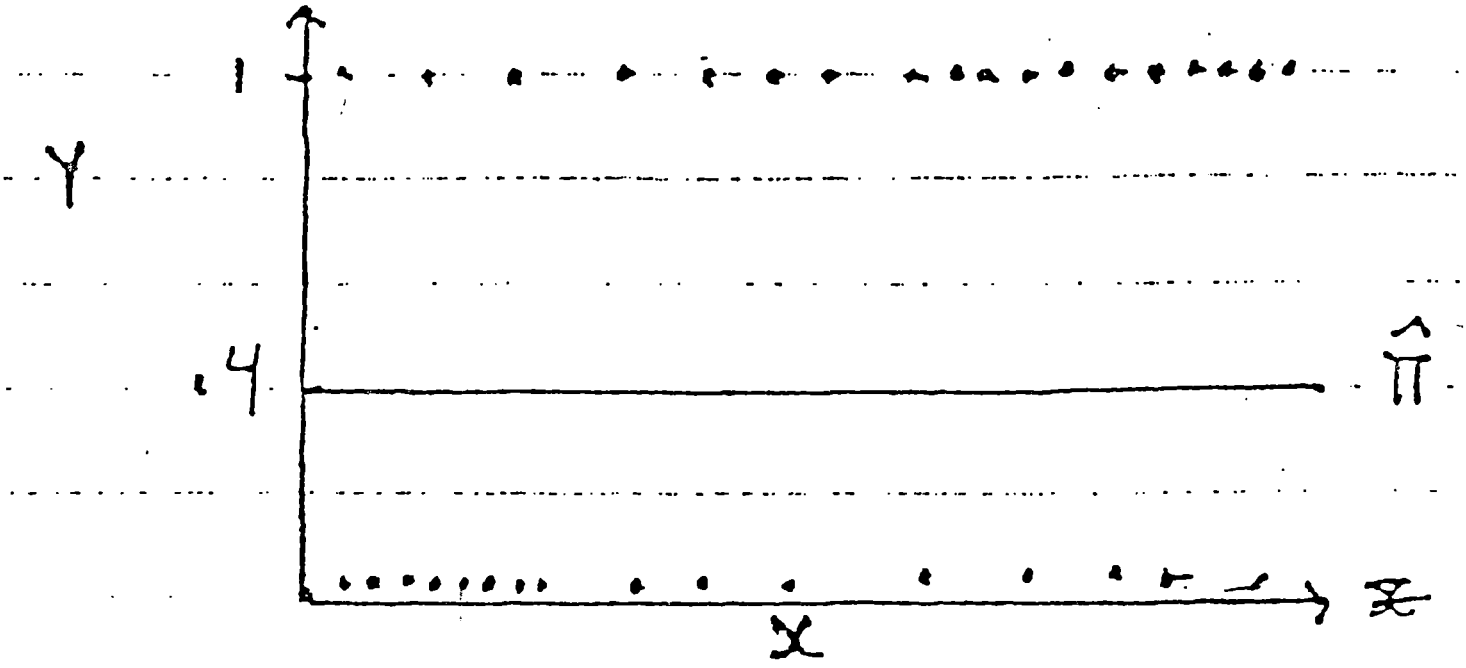


Figure 4. Constant model

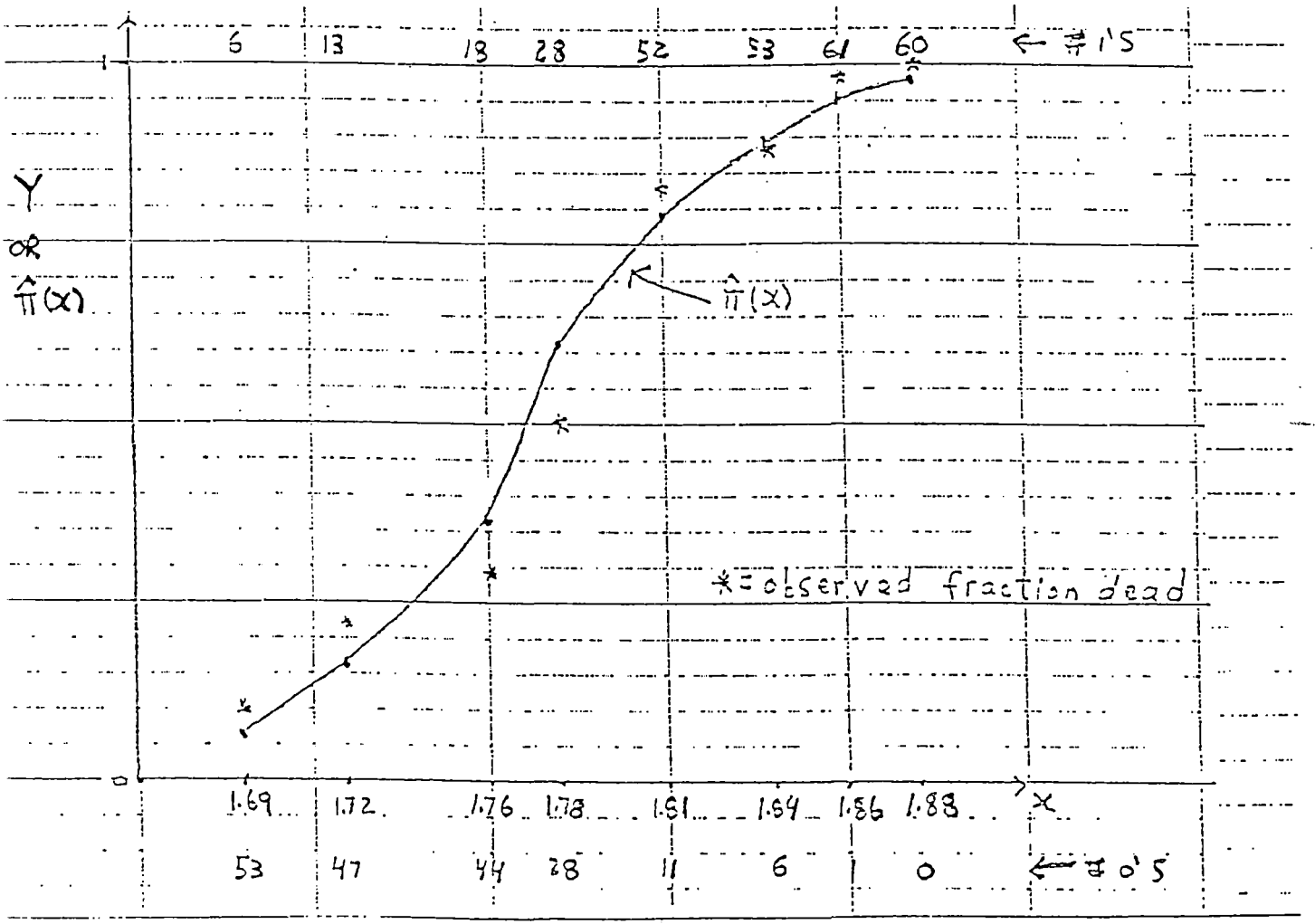


Figure 5. Logistic and best $\hat{\pi}$ models for beetle poison example.

Probability of a Fire Day

Lolo National Forest, ERC(G), 1970–85

* gives the observed fraction of fire days at the ERC(G) values.

— logistic model

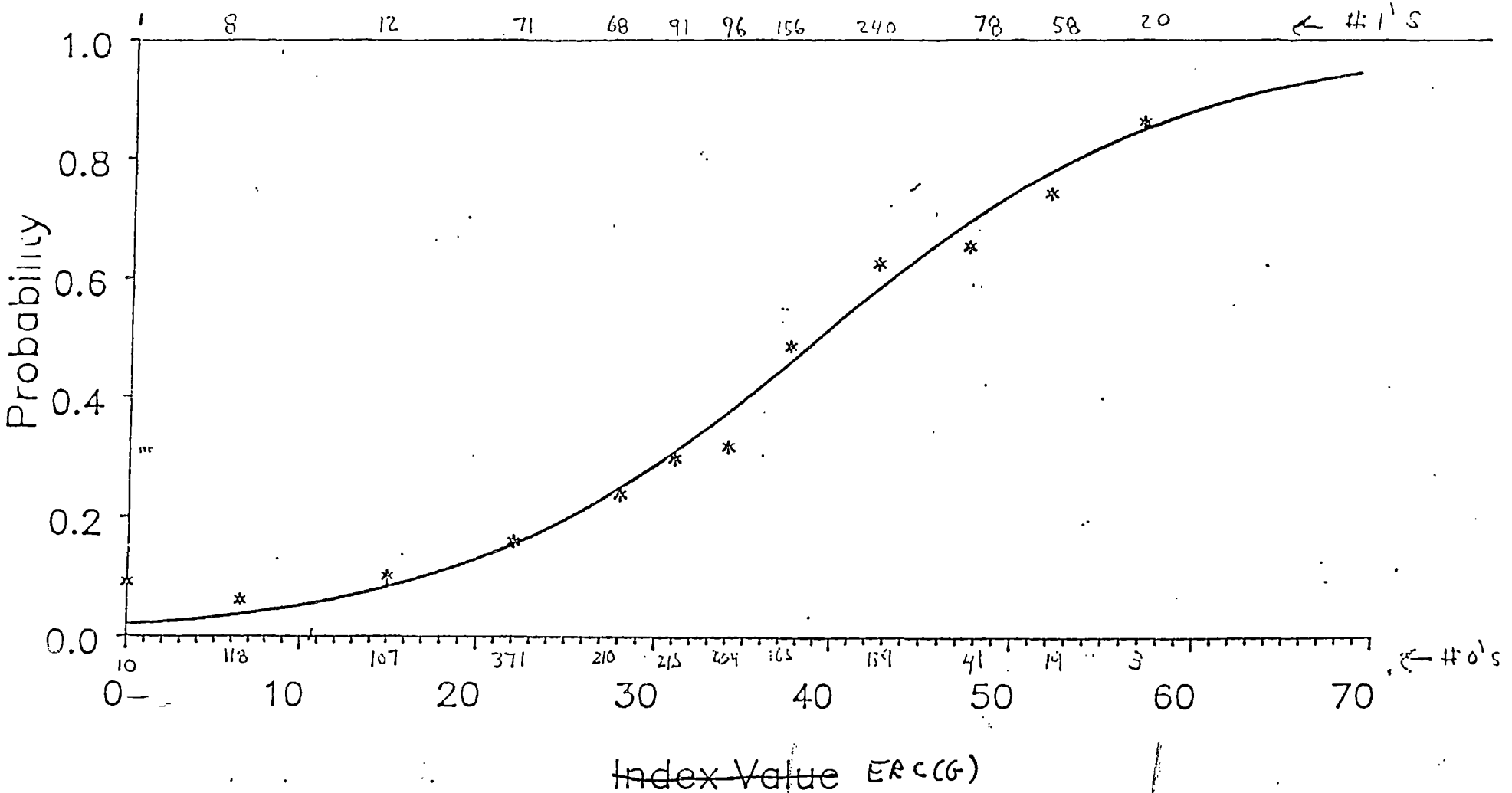


Figure #6